DOCUMENT RESUME

ED 426 072 TM 029 294

AUTHOR Wang, Tianyou; Hanson, Bradley A.; Harris, Deborah J.
TITLE The Effectiveness of Circular Equating as a Criterion for Evaluating Equating.
INSTITUTION American Coll. Testing Program, Iowa City, IA.
REPORT NO ACT-RR-98-6
PUB DATE 1998-10-00
NOTE 31p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168 Iowa City, IA 52243-0168.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; Evaluation Methods; Heuristics; Sampling; Simulation; Test Construction; *Test Format
IDENTIFIERS *Circular Equating; Randomization

ABSTRACT

Equating a test form to itself through a chain of equatings, commonly referred to as circular equating, has been widely used as a criterion to evaluate the adequacy of equating. This paper uses both analytical methods and simulation methods to show that this criterion is in general invalid in serving this purpose. For the random groups design done in the same year, it is shown analytically that circular equating will always result in the identity function (i.e., the perfect result) even with the presence of random and systematic equating errors. For the random groups design done in the different years, a heuristic argument is provided that circular equating will generally deviate from the identity function by some random sampling error. A simulation study for this design also showed that the expected values of the circular equating may deviate slightly from the identity function but those deviations do not reflect the systematic error (bias) embedded in the equating. For the common-item nonequivalent groups design, a simulation study was done to show that circular equating again can not reflect the systematic error in equating. More effective ways of assessing random and systematic equating errors are recommended. (Contains 4 tables, 4 figures, and 13 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACT Research Report Series 98-6

ED 426 072

The Effectiveness of Circular Equating as a Criterion for Evaluating Equating

Tianyou Wang

Bradley A. Hanson

Deborah J. Harris

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Patricia Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029294

ACT

October 1998

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

© 1998 by ACT, Inc. All rights reserved.

BEST COPY AVAILABLE

The Effectiveness of Circular Equating as a Criterion for Evaluating Equating

Tianyou Wang
Bradley A. Hanson
Deborah J. Harris

Abstract

Equating a test form to itself through a chain of equatings, commonly referred to as circular equating, has been widely used as a criterion to evaluate the adequacy of equating. This paper uses both analytical methods and simulation methods to show that this criterion is in general invalid in serving this purpose. For the random groups design done in the same year, it is shown analytically that circular equating will always result in the identity function (i.e., the perfect result) even with the presence of random and systematic equating errors. For the random groups design done in the different years, a heuristic argument is provided that circular equating will generally deviate from the identity function by some random sampling error. A simulation study for this design also showed that expected values of the circular equating may deviate slightly from the identity function but those deviations do not reflect the systematic error (bias) embedded in the equating. For the common-item nonequivalent groups design, a simulation study was done to show that circular equating again can not reflect the systematic error in equating. More effective ways of assessing random and systematic equating errors are recommended.

Acknowledgments

The authors wish to thank Michael Kolen, Jill Crouse and Ronald Cope for their helpful comments on the earlier drafts of this report.

The Effectiveness of Circular Equating as a Criterion for Evaluating Equating

In test equating, there has been a lack of definitive and practically feasible criteria for evaluating the adequacy of equating. Harris and Crouse (1993) did a thorough review and discussion of the available criteria in the literature. One of the criteria they reviewed is the circular equating paradigm. Circular equating involves equating a test form to itself through a chain of equatings. To illustrate this with a case of three test forms, X, Y, and Z, test form X is equated to form Y, which is equated to form Z, which is equated back to form X. It is presumed that if an equating is without error, an identity circular equating should result, and that if the equating functions in the chain contains much error, the circular equating would not result in identity equating and the result should reflect the error accumulated in the chain. Based on this reasoning, the circular equating criterion was commonly used in evaluating equating methods (e.g., Cope, 1987) or scale drift in IRT equating (e.g., Petersen, Cook, & Stocking, 1983). Brennan and Kolen (1987a, b) and Angoff (1987) discussed this criterion. Angoff (1987) had more positive views on the usefulness of this criterion. Brennan and Kolen (1987a, b), however, expressed cautions about using this criterion. They pointed out that no equating at all will result in identity equating under this paradigm. They also demonstrated that equating methods with fewer parameters tend to achieve better results, and starting from a different form may affect the results. Despite these concerns, this paradigm and some variations of it continue to be used as a criterion in both research and practice (e.g., Klein & Jarjoura, 1985; McKinley & Schaeffer, 1989; Gafni & Melamed, 1990; Harris, Welch & Wang , 1994). The applications of the criterion have not produced clear results about its usefulness. Some authors expressed doubts about the validity of this criterion (e.g., Gafni & Melamed, 1990). There has not been a substantive study on the validity and usefulness of this widely used criterion. The objective of this paper is to address this need. More specifically, we focus on type of equating error for which this criterion can or cannot provide an accurate measure, when applied to different equating methods and equating designs.

Kolen and Brennan (1995, pp. 210-211) summarized two major types of equating errors: random error and systematic error. There is one source of random error; that is, equating is performed based on samples randomly drawn from the population of examinees rather than based on the population itself. There are two major sources of systematic errors. One source is the equating method used, including violations of the assumptions associated with the equating method, and the estimation bias related to that method. A second source is the collection of data in the equating study, including whether the samples are randomly drawn from the population that actually take the test forms, and whether the equating design was properly implemented. A criterion for evaluating equating should be able to provide an accurate measure of one or both types of equating error and be able to indicate the amount of error present. Thus the effectiveness of the circular equating paradigm as a criterion for evaluating equating is determined by the extent to which it can meet these requirements.

The variance or standard deviation of equating across samples (the latter is called the standard error of equating) is usually used to assess the magnitude of random error. Assessing standard error of equating usually involves drawing random samples from the same populations under the same set of conditions. Since in practice circular equating only involves one set of samples without replication, it is not expected that circular equating in and of itself could be used to assess the standard error of equating. Traditionally, the circular equating criterion was intended to assess systematic error. Other statistical techniques such as the bootstrap methods can be used to assess standard error of equating for a single equating or for a chain of equatings for different equating designs (Hanson, Harris & Kolen, 1997; Hanson, 1996; Hanson, 1998). For these reasons, the study will focus more on the effectiveness of circular equating in evaluating systematic error rather than random error.

It is not known *a priori* whether the effectiveness of circular equating as a criterion is the same under different equating study designs. This paper considers three categories of test equating designs: the random groups design, the common-item nonequivalent groups design, and the single group counter-balanced design (Kolen & Brennan, 1995, Chapter 1). We will study the two most

commonly used equating designs: the random groups design and the common-item nonequivalent groups design, in our investigation of circular equating. We will discuss the single group counter-balance design at the end of the paper.

Random Groups Design

There are two possible scenarios of applying circular equating under the random groups design. One is to have a chain of equatings with a single data collection; that is, the groups that are administered these different test forms are randomly equivalent groups. For example, three randomly equivalent groups are administered three test forms: form X, Y, and Z, and form X is equated and to Y, Y is equated to Z, and Z is equated back to X. For convenience, we will call this circular equating under random groups done in the same year. Another scenario is to have samples from different populations for the different equating chains in a circle. Taking the previous example, form X is equated to form Y in one year, and form Y is equated to form Z in a second year with another random groups data collection, and form Z is equated back to X in a third year with a third data collection. We will call this circular equating under random groups done in different years. We will show analytically that under the first scenario, the circular equating paradigm is basically an invalid criterion because it cannot reveal any type of equating error. In another word, circular equating will result in perfect results (identity equating) with any known equating method.

Under the second scenario, we will give analytical results for a special case and use simulation to study what type of equating error can be reflected by circular equating in other cases. In the case for which analytical results are given, we will show that only the random error in the statistics used to estimate the equating function cause the circular equating to deviate from the perfect result, and that systematic errors can not be detected by this criterion. That will explain why equating methods with fewer parameters appear to achieve better results under circular equating, as Brennan and Kolen (1987) contended.

Simulation methods (also called parametric bootstrap method; Efron, 1982; Efron & Tibshirani, 1993) will also be used to assess the effectiveness of circular equating under the second scenario. In the simulation study, the population distribution for each of the test forms are estimated from real data, samples of test data are then generated from these population distributions, and circular equating is performed using the generated data.

Analytical Results for Circular Equating Done in the Same Year

We shall consider the case of circular equating of three test forms done in the same year. The results can be extended to cases of more than three forms. Assume we have forms X, Y and Z. Let $m_X, s_X, F_X(x)$ be the sample mean, standard deviation, and the cumulative distribution function for a particular form X, and let $e_{X\to Y}(x)$ stand for the equating function from form X to Y; that is, the form Y equivalent of a particular score x on form X. Notations for the other forms and other equating functions are defined likewise with only the subscript changed accordingly. As Kolen and Brennan (1995, p. 9) stated, all equating procedures are required to have the symmetry property. By this property, we have

$$e_{X\to Z}(x) = e^{-1}_{Z\to X}(x)\ , \tag{1}$$

where the superscript -1 means the inverse function. Equivalently, we have

$$e_{Z\to X}\left[e_{X\to Z}(x)\right] = x \tag{2}$$

It is generally true with any of the known equating methods that for random groups design done in the same year, the equating function from form X to Z directly equals to the equating function from X to Z through another form Y (or through a chain of equatings if there are more than three forms in all). Symbolically, this can be expressed as

$$e_{Y\to Z}\left[e_{X\to Y}(x)\right] = e_{X\to Z}(x) \tag{3}$$

This assertion is not obvious but can be verified case by case for various equating methods. We will verify Equation 3 for the linear and equipercentile methods. Verifications for other methods basically follow the same logic.

For the linear equating method, we have (see Kolen & Brennan, 1995, p. 30)

$$e_{X\to Y}(x)=\left(\frac{x-m_X}{s_X}\right)s_Y+m_Y \quad . \tag{4}$$

Therefore

$$e_{Y\to Z}\left[e_{X\to Y}(x)\right]=\left\{\frac{\left[\left(\frac{x-m_X}{s_X}\right)s_Y+m_Y\right]-m_Y}{s_Y}\right\}s_Z+m_Z=\left(\frac{x-m_X}{s_X}\right)s_Z+m_Z=e_{X\to Z}(x) \tag{5}$$

Since mean equating is only a special case of the linear equating, Equation 5 is sufficient for the mean equating method. For the equipercentile equating, we have for continuous score distributions (see Kolen & Brennan, 1995, p. 36)

$$e_{X\to Y}(x)=F_Y^{-1}\left[F_X(x)\right] \quad . \tag{6}$$

Therefore

$$e_{Y\to Z}\left[e_{X\to Y}(x)\right]=F_Z^{-1}\left\{F_Y\left[F_Y^{-1}(F_X(x))\right]\right\}=F_Z^{-1}\left[F_X(x)\right]=e_{X\to Z}(x) \tag{7}$$

This holds regardless of whether smoothing is performed on the score distributions. Smoothing of the score distributions before equipercentile equating is applied is called presmoothing. If smoothing is performed on the equipercentile equating function, it is called postsmoothing (see Kolen & Brennan, 1995, Chapter 3 for a discussion of various smoothing methods). Equation 7 shows that Equation 3 is valid both for unsmoothed equipercentile equating and equipercentile equating with presmoothing. However, the discreteness of the score distributions and the linear interpolation underlying the equipercentile equating method may cause the equality in Equation 7

not to hold exactly, but the deviation is expected to be quite small. With postsmoothing, Equation 3 may also not hold exactly, because the postsmoothing method smoothes the equating function and thus changes the equating relationship. But because the deviation is expected to be small and is caused only by the smoothing factor (which is not our main concern here), the deviation from Equation 7 does not invalidate the main argument. For now, we assume that Equation 3 holds. Substituting the left-hand side of Equation 3 for $e_{X\to Z}(x)$ in Equation 2, yields

$$e_{Z\to X}\{e_{Y\to Z}[e_{X\to Y}(x)]\} = x \tag{8}$$

Equation 8 indicates that circular equating under this design will always result in the identity function even though there may be random or systematic error in each link of equatings. In other words, the circular equating can not provide an accurate measure of any equating error and therefore is an invalid criterion for evaluating equating under this data collection design. In cases where there are more then three forms in the circle, Equation 3 can be applied repeatedly, and when combined with Equation 2, a result analogous to Equation 8 can be obtained.

Circular Equating Done in Different Years

For this equating design, we shall only present a proof for the mean equating method. Suppose we have three equating links done in three different years: form X is equated to form Y in the first year, form Y is equated to form Z in the second year, and form Z is equated back to form X in the third year. Assume that there is no interaction between the test forms and the examinee populations of the different years; that is, if there is an effect of the population ability levels on the test scores, the effect is the same for the different forms. We also assume the population effects are additive and form-invariant; that is, the effect of a particular population can be added to the sample means and remain constant for different test forms. Let m_{X1} and e_{X1} be the sample mean and the associated random error for form X from the first year population. Let μ_X be defined as the expected mean for form X from all concerned populations, and μ_{Xi} be the expected mean for form X and population i; that is

$$\mu_X = \frac{1}{3}\sum_{i=1}^{3}\mu_{Xi} \ . \tag{9}$$

Expected means for form Y and Z are similarly defined. Let d_1 be the effect of the first year population; that is, $d_1 = \mu_{X1} - \mu_X = \mu_{Y1} - \mu_Y = \mu_{Z1} - \mu_Z$. Notations for other population effects are defined likewise, with appropriate subscripts. We have

$$m_{X1} = \mu_X + d_1 + e_{X1} \ , \tag{10}$$

$$m_{X3} = \mu_X + d_3 + e_{X3} \ , \tag{11}$$

$$m_{Y1} = \mu_Y + d_1 + e_{Y1} \ , \tag{12}$$

$$m_{Y2} = \mu_Y + d_2 + e_{Y2} \ , \tag{13}$$

$$m_{Z2} = \mu_Z + d_2 + e_{Z2} \ , \tag{14}$$

$$m_{Z3} = \mu_Z + d_3 + e_{Z3} \ . \tag{15}$$

Therefore

$$\begin{aligned} e_{Z\to X}\left\{e_{Y\to Z}\left[e_{X\to Y}(x)\right]\right\} &= x - m_{X1} + m_{Y1} - m_{Y2} + m_{Z2} - m_{Z3} + m_{X3} \\ &= x - (\mu_X + d_1 + e_{X1}) + (\mu_Y + d_1 + e_{Y1}) - (\mu_Y + d_2 + e_{Y2}) \\ &+ (\mu_Z + d_2 + e_{Z2}) - (\mu_Z + d_3 + e_{Z3}) + (\mu_X + d_3 + e_{X3}) \\ &= x - e_{X1} + e_{Y1} - e_{Y2} + e_{Z2} - e_{Z3} + e_{X3} \end{aligned} \tag{16}$$

Equation 16 shows that under the previously described assumptions, a mean circular equating will result in the identity function except for some random errors. The systematic error due to population differences or equating methods can not be reflected by this circular equating. It is seen that in the chains of equatings that constitute this circle, the systematic errors due to population differences cancel out, and that any systematic errors embedded in the equating method (for example, if the mean equating method is used when the mean function does not represent the true population equating function) never enter the process.

This reasoning process can, to some extent, be extended to the linear equating method and to other more complicated equating methods even though it may be hard to prove in a clear-cut

fashion. The simulation results presented below will provide further evidence to support this argument.

Simulation Study for Circular Equating Done in Different Years

Method: In order to manipulate the population distributions for different test forms for different years, an IRT model was used to compute the population score distributions. The three parameter logistic (3PL) model was fit to regular equating data for three forms of the ACT Assessment Mathematics test, and the item parameters were estimated using the program EM1 (Zeng, 1995). The summary statistics for the item parameters for forms A, B, and C are contained in Table 1. Table 1 shows that the means of the b parameters of the test forms differ somewhat. In particular, form A has a higher mean b value than the other two forms. Three sets of population θ distributions were used to compute the population score distributions. The population θ distributions are all normally distributed with their means and standard deviations contained in Table 2a. These three sets are intended to represent three different situations in the change of the population distribution. Set one represents a situation where populations are stable across years. Set two represents a situation where the changes are in the same direction. Set three represents a situation where the changes are in a different direction. The steps used in computing the population score distributions are:

1). Conditioned on a given θ, the score distribution $f(x|\theta)$ is computed using the Lord and Wingersky (1984) recursive algorithm.

2). The marginal score distribution is computed using the following equation:

$$f(x) = \int_{\theta} f(x|\theta)\psi(\theta)d\theta \quad , \tag{17}$$

where $\psi(\theta)$ is the population distribution of θ. Some form of numerical integration can be used to carry out this step. This computational procedure is also described in Kolen and Brennan (1995, pp. 182-183).

After these population score distributions were computed, they were input into a computer program RG Equating Error (Hanson, 1996) which can estimate the mean and standard deviation of repeated circular equating functions. Using this program, the simulation process takes the following steps:

1). Random samples of scores for 2000 simulated examinees were drawn from the population score distributions. Six samples were drawn, with each test form having one random sample from each of two populations. The six samples were: forms A and B for the first year population distribution; forms B and C for the second year population distribution; and forms C and A for the third year population distribution.

2). Circular equating was performed for this set of 6 random samples . form A was equated to form B using the pair of samples drawn from first year population; form B was equated to form C using the pair of samples drawn from the second year population; form C was equated back for form A using the pair of samples drawn from the third year population.

3). Steps 1 and 2 were replicated 5000 times. The mean and standard deviation of the circular equating function over 5000 replications were computed.

4). Steps 1 through 3 were repeated for three different equating methods: Mean equating, linear equating and equipercentile equating with log-linear presmoothing with a degree of six (Kolen & Brennan, 1995, Chapter 3). This degree of smoothing was considered to be sufficient for this situation.

5). Steps 1 through 4 were repeated for three different sets of populations (See Table 2).

Results: The mean deviations and standard errors of the circular equating function for this design are plotted in Figure 1. The mean deviations in Figure 1 are the means of the equated scores through circular equating across 5000 replications minus the identity function. It can been seen from Figure 1, that with population set 1 there is not much mean deviation for any of the methods except at the extreme scores for the equipercentile equating with presmoothing. A comparison across population sets shows that the mean deviation of the equipercentile (with presmoothing) method was least affected by the change of the population. For the mean and linear

methods, the mean deviation tended to be greater when the population changes in the same direction (population set 2) than in a different direction (population set 3). The mean deviation from the identity function for the mean equating method for population sets 2 and 3 appear to contradict Equation 16. This may be due to some violation of the assumptions that were used to derive Equation 16. In particular, the assumption that the population effects are additive and form-invariant is probably violated to some extent because an IRT model was used to generate the response data and the score distribution. Overall, the mean deviations for all these methods are not very large, particular at the middle score range. The standard errors (SE) plotted in Figure 1 are the standard deviations of the equating functions through circular equating across the 5000 replications. The SE plots show that the mean equating consistently has smallest SE, and that the linear equating has a smaller SE than the equipercentile (with presmoothing) method over most of the score range. The random errors usually have larger magnitude than the mean deviation and constitute a larger part in the total error.

The average standard error and the average absolute mean deviation across score points are contained in Tables 3a and 3b. The values in Table 3 were computed by averaging the values reported in Figure 1 over the number correct score scale weighted by the frequency distribution of form A. Table 3c contains the average root mean square error. The average root mean square error is the square root of the sum of the squared average standard deviation and squared absolute mean deviation. The average absolute mean deviation, average standard error, and average root mean squared error are measures of systematic, random, and total error in the estimates, respectively. The results in Table 3 are consistent with the pattern shown in Figure 1.

To facilitate interpreting the results from these simulations, the true equating functions were computed and plotted as shown in Figure 2. True equating is defined as the unsmoothed equipercentile equating function performed on the population distributions. Figure 2 shows that the true equating function from A to B is curvilinear and deviates from the identity function about 1 to 2 points. The true equating function from B to C is close to being linear. The true circular equating from A to B to C to A is basically the identity function. Figure 2 also shows that all the

true equating functions remain basically the same across three population sets. These plots suggest that the true equating functions in the chains of the circle deviate from linear or mean equating functions, which means that the mean and linear equating methods are biased for some of the chains in the circle. Figure 2 also shows there is little population effects.

To further illustrate this, the plots in figure 3 show the bias in the mean and linear equating methods for individual links in the chain of equatings for population set 2. The Figure 3a gives the difference of the mean and linear equating from the equipercentile equating of form C to form A computed using the population data for the third year population. This gives the bias of the linear and mean equating methods relative to the equipercentile method in the population. The Figure 3b gives the bias in the linear and mean equating methods for the form B to form C to form A equating as computed using the population distributions. It gives the bias in the form B to form A equating that results from combining the form C to form A equating in the third year population with the form B to form C equating in the second year population. Figure 3c gives the bias (i.e., the deviation from the population equipercentile equating function) of the linear and mean methods for the full circular equating computed using the population distributions.

Figure 3d gives the deviations of the mean, linear and equipercentile equating functions from the identity function for the full circular equating computed using the population distributions. The circular equipercentile equating function computed using the population distributions is close to, but not exactly, an identity function. The differences given in Figure 3d are very close to the mean deviation computed using simulation as reported in Figure 1b. This similarity indicates that the mean deviations plotted in Figure 1 accurately reflect the deviations of the population circular equating from the identity function. Again deviations from the identity functions seems to be caused by violations to the assumptions used to derive Equation 16, but these minor deviations do not invalid the general argument that circular equating can not reflect the systematic error due to population differences or equating methods.

It can be further observed from the plots in Figure 3 that the bias of the mean and linear methods in the links that make up the circular equating is much greater than their mean deviation in

the full circular equating. Only considering the mean deviation for the full circular equating would misrepresent the bias in the mean and linear methods when equating two different forms.

In light of these results, it can be seen that the circular equating with a simpler method (such as the mean equating; see Table 3c) tends to result in less deviation from the identity function than with a more sophisticated method even when the simpler method may not be appropriate for some of the equating links that form the chain (e.g., Figure 3). Thus, it can be concluded that circular equating is not a good criterion to use in assessing systematic error embedded in an equating method, and therefore is not an appropriate criterion for comparing different equating methods. The deviations from the identity function mostly reflect random sampling error, and to a lesser extent reflect a change in the population distribution from year to year.

Common item nonequivalent groups design

With this design, a new test form X and an old form Y are administered at two test dates to two supposedly non-equivalent groups of examinees. Test form X is equated to form Y through a common set of items, V. If V is counted as a part of the score reported for forms X and Y, then it is called an internal anchor; if V is not counted in the scoring of forms X and Y, it is called an external anchor.

The rationale underlying almost all the equating methods under this design is as follows: First, the sample statistics of form X are projected to the group that takes only form Y through the relationship between form X and the common set V. The same thing is done for form Y. Second, a synthetic group (sometimes called the synthetic population) is formed as a weighted combination of the groups taking forms X and Y and the sample statistics for forms X and Y are projected to the synthetic group. Finally, with the sample statistics for both forms for the synthetic group, the equating procedure is done in the same fashion as in the random groups design. In the case of circular equating, such as with forms X, Y, and Z, a different synthetic group is formed for each of the equating links in the circle and, for each synthetic group, it is as if a random groups design is employed. In this way, circular equating done under this design is similar to the random groups

design done in different years as described in the previous section. The differences are that the synthetic groups are hypothetical rather than real and the form statistics for the synthetic groups are estimated rather than directly collected as in the random groups design. The analytical results for the random groups design can shed light on the common-item nonequivalent groups design. It is hypothesized that under the common item nonequivalent groups design, circular equating can not provide an accurate measure of the systematic error caused by the equating methods and thus can not be used to compare different equating methods. It is also hypothesized that if the population does not change much from year to year, circular equating would result in the identity function except for some random deviations. How the change in the populations from year to year will affect the systematic deviations needs empirical investigation. For this reason and for confirming our hypotheses, a simulation was conducted.

Simulation Study for Circular Equating for the Common-Item Design

Method: Four forms of the ACT Mathematics test, A, B, C, and D were used to form a circle (form A, B, and C are the same forms used in the simulation study for the random groups design done in different years). A form E was split into four equivalent sets of items in a spiral fashion, with 15 items in each set. These four sets of items are used as external anchor items for the circular equating from A to B to C to D and back to A. IRT models were used in generating the score distributions. The summary statistics of item parameters for these forms are summarized in Table 1. Three sets of population θ distributions were used in computing the bivariate score distributions between the form scores and anchor set scores. The θ distributions are all normally distributed with their means and standard deviations contained in Table 2b. Population set 1 is designed to represent a situation where the examinee populations do not change from year to year. Population set 2 represents a situation where the changes are in the same direction. Population set 3 represents a situation where the changes are in different directions. The computer program CI Equating Error (Hanson, 1998) was used to carry out the simulation. The equating sample size was 2000, and the number of replications was 5000. The steps used in the simulation were very

similar to the steps described for the simulation study for random groups design done in different years, except that bivariate distributions were computed and used in the equatings under a common-item design. Three equating methods were used: the Tucker mean equating method, the Tucker linear equating method, and the equipercentile (unsmoothed) equating method (Kolen & Brannan, 1995, pp. 107-111). The means and standard deviations of the circular equating over replications were recorded.

Results: The mean deviation and SE of the circular equating under this design are plotted in Figure 4. The mean deviation in Figure 4 is the mean of the equated scores through circular equating across 5000 replications minus the identity function. The SEs are the standard deviations of the circular equating function across the 5000 replications. Table 4 contains the average absolute mean deviation, average SE, and average root mean square error as averaged across the number correct score scale weighted by the form A distribution in the first year population.

The results for both the mean deviation and SEs are similar to those from the random groups design done in different years. For population set 1, the mean deviation is basically zero for the Tucker mean and linear methods, and is also very small for the equipercentile method except at extreme scores. The mean deviation was affected when the populations changed from year to year. It can be seen that when population changes are in the same direction (population set 2), the mean deviation is relatively small compared to when populations changes are in different directions (population set 3). This result is different than that observed in the simulation study for the random groups design done in different years. The plots of SEs show that the Tucker mean method consistently had smallest SEs and the equipercentile method had the largest. The averaged SEs over score points are contained in Table 4b which shows the same pattern. The SEs are virtually unaffected by the change in the populations. In general, these results confirm our hypotheses and suggest circular equating is not a valid criterion to evaluate the systematic errors in the equatings.

Discussion and Conclusions

The circular equating paradigm has been widely used as a criterion for evaluating equating errors, and for comparing different equating methods without a clear understanding of the effectiveness of the criterion itself. Although some authors have expressed reservations about the effectiveness of the paradigm, there has not been an empirical study on this issue. The purpose of this study was to clarify what type of equating error this paradigm can or cannot evaluate under various equating designs and situations.

The effectiveness of circular equating as an equating criterion was investigated using analytical methods and simulation. For the random groups design where all forms are administered to randomly equivalent groups, it was shown analytically that circular equating will always result in the identity function. Thus, circular equating in this case is an invalid equating criterion for comparing equating methods. Analytical results were also obtained for the mean equating method under the random groups design where pairs of forms to be equated are administered to samples from different populations under the assumption of no interaction between population and form differences. In this case the only difference between mean circular equating and the identity function is due to random error, so circular equating cannot be used for comparing equating methods or observing population effects. The analytical results for the special cases considered showed that circular equating is not an acceptable equating criterion due to the fact it cannot detect systematic equating error from either source.

The performance of circular equating was investigated for some more complicated cases using simulation. Simulations were performed for a random groups design in which pairs of forms equated in the chain were administered to samples from different populations, and for the common-item nonequivalent groups design. The simulations were used to compute systematic and random components in the differences between circular equating results and the identity function. In all the cases considered the major portion of the differences between the circular equating and the identity equating were due to random error. This was true even when there was considerable systematic error (bias) in an equating method for intermediate links in the circular equating chain.

In evaluating an equating method it is more difficult to assess systematic errors than random errors. Random equating error can be investigated in a straightforward manner using the bootstrap (Efron, 1982; Efron & Tibshirani, 1993; Hanson, 1996, 1998), and for many equating methods analytical standard errors of equating are available (Kolen and Brennan, 1995). Circular equating was considered a possible method for evaluating the systematic error in an equating method. However, the simulations showed that most of the deviation of circular equating from an identity equating is due to random variation, rather than systematic variation. In the simulations where systematic deviations of circular equating from the identity equating did exist, they tended to be much smaller than the systematic error in the equating method for equating links within the circle. Thus the deviations of the circular equating from the identity equating did not accurately represent the systematic equating error for the method on the links in the circle.

The results in this paper support the conclusion that circular equating does not provide a useful criterion by which the accuracy of equating methods can be investigated. One alternative to circular equating for investigating systematic error in an equating method is simulation. The disadvantage of simulation is that a model needs to be specified to perform the simulation, and it can be unclear to what extent the model chosen produces realistic data. An alternative to circular equating that does not involve simulation is to obtain data so that an equating between two forms can be done directly (e.g., form A can be directly equated for form C), or indirectly (e.g., form A equated to form B equated to form C). Differences in equating results using the indirect or direct links would include both systematic and random error, but might give a more realistic assessment of systematic error than circular equating.

The performance of circular equating as a criterion for the single group counter-balanced design was not studied. Because of the similarity in the equating methodology between this design and the random groups design, it is safe to conclude that circular equating is also ineffective in evaluating the equating error under this design.

References

Angoff, W.H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291-300.

Brennan, R.L., & Kolen, M.J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11,* 279-290.

Brennan, R.L., & Kolen, M.J. (1987). A reply to Angoff. *Applied Psychological Measurement, 11,* 301-306.

Cope, R.T. (1987). How well do the Angoff design V linear equating methods compare with the Tucker and Levine methods? *Applied Psychological Measurement, 11,* 143-149.

Gafni, N., & Melamed, E. (1990). Using the circular equating paradigm for comparison of linear equating models. *Applied Psychological Measurement, 14,* 247-256.

Hanson, B. A. (1998). *CI Equating Error: A program for computing bootstrap standard errors of equating for the common-item nonequivalent group equating design.* Computer software document. Iowa City, IA: ACT.

Hanson, B. A. (1996). *RG Equating Error: A program for computing bootstrap standard errors of equating for the random groups equating design.* Computer software document. Iowa City, IA: ACT.

Harris, D.J., & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195-240.

Harris, D.J., Welch, C.J., & Wang, T. (1994). *Issues in equating performance assessments.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Klein, L.W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197-206.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices.* New York: Springer.

McKinley, R.L., & Schaeffer, G.A. (1989). *Reducing test form overlap of the GRE subject test in mathematics using IRT triple-part equating.* ETS Research Report 89-8, Educational Testing Service: Princeton, NJ.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 2,* 137-156.

TABLE 1

The Summary Statistics of the Item Parameters for the Test Forms Used in the Equatings

Form/par.	n	Mean	Std. Dev.	Minimum	Maximum	Skewness	Kurtosis
Form A							
a	60	1.035	0.258	0.628	1.651	0.393	-0.732
b	60	0.324	0.932	-1.816	2.199	-0.265	-0.513
c	60	0.149	0.049	0.058	0.251	-0.053	-0.805
Form B							
a	60	0.960	0.314	0.349	1.681	0.201	-0.582
b	60	0.182	0.987	-2.132	1.911	-0.300	-0.568
c	60	0.155	0.041	0.075	0.247	0.114	-0.557
Form C							
a	60	0.950	0.276	0.524	1.509	0.379	-0.854
b	60	0.115	0.879	-1.851	1.749	-0.402	-0.605
c	60	0.141	0.039	0.042	0.232	-0.464	0.551
Form D							
a	60	1.012	0.325	0.296	1.635	-0.122	-0.499
b	60	0.116	0.943	-2.348	1.799	-0.600	-0.091
c	60	0.144	0.046	0.041	0.252	-0.040	-0.333
Form E							
a	60	0.940	0.313	0.441	1.933	0.569	0.562
b	60	0.171	0.998	-2.515	1.971	-0.378	-0.345
c	60	0.151	0.047	0.031	0.250	-0.343	-0.070
Form F							
a	60	0.936	0.246	0.403	1.520	-0.068	0.015
b	60	0.105	0.974	-1.871	1.771	-0.183	-0.769
c	60	0.146	0.045	0.062	0.256	0.240	-0.209

TABLE 2

The Means and SDs for the Normally Distributed Populations

a. For the random groups design done in different years.

Set one	Mean	SD
The first year population	0.0	1.0
The second year population	0.0	1.0
The third year population	0.0	1.0
Set two		
The first year population	0.0	1.0
The second year population	0.2	1.1
The third year population	0.4	1.2
Set three		
The first year population	0.0	1.0
The second year population	0.2	1.1
The third year population	-0.2	0.9

b. For the the common-item nonequivalent groups design.

Set one	Mean	SD
The first year population	0.0	1.0
The second year population	0.0	1.0
The third year population	0.0	1.0
The fourth year population	0.0	1.0
Set two		
The first year population	0.0	1.0
The second year population	0.1	1.1
The third year population	0.2	1.2
The fourth year population	0.3	1.3
Set three		
The first year population	0.0	1.0
The second year population	0.2	1.2
The third year population	0.0	1.0
The fourth year population	-0.2	0.8

TABLE 3

The Average Standard Error, Absolute Mean Deviation and Root Mean Square Error for the Random Groups Design Done in Different Years.

a. Average Standard Errors.

	Equating Method		
Populatoin Set	Mean	Linear	Presmoothing
Set One	0.666	0.759	0.894
Set Two	0.714	0.810	0.962
Set Three	0.662	0.762	0.905

b. Average Mean Deviation.

	Equating Method		
Populatoin Set	Mean	Linear	Presmoothing
Set One	0.000	0.003	0.018
Set Two	0.199	0.120	0.025
Set Three	0.077	0.049	0.023

c. Average Root Mean Square Error.

	Equating Method		
Populatoin Set	Mean	Linear	Presmoothing
Set One	0.666	0.759	0.894
Set Two	0.742	0.818	0.962
Set Three	0.667	0.764	0.905

TABLE 4

The Average Standard Error, Absolute Mean Deviation and Root Mean Square Error for the Common-Item Nonequivalent Groups Design Done in Different Years.

a. Average Standard Errors.

	Equating Method		
Populatoin Set	Tucker Mean	Tucker Linear	Presmoothing
Set One	0.405	0.513	0.785
Set Two	0.413	0.497	0.787
Set Three	0.421	0.539	0.818

b. Average Mean Deviation.

	Equating Method		
Populatoin Set	Tucker Mean	Tucker Linear	Presmoothing
Set One	0.004	0.002	0.014
Set Two	0.050	0.069	0.103
Set Three	0.347	0.195	0.189

c. Average Root Mean Square Error.

	Equating Method		
Populatoin Set	Tucker Mean	Tucker Linear	Presmoothing
Set One	0.405	0.513	0.785
Set Two	0.416	0.502	0.794
Set Three	0.546	0.573	0.839

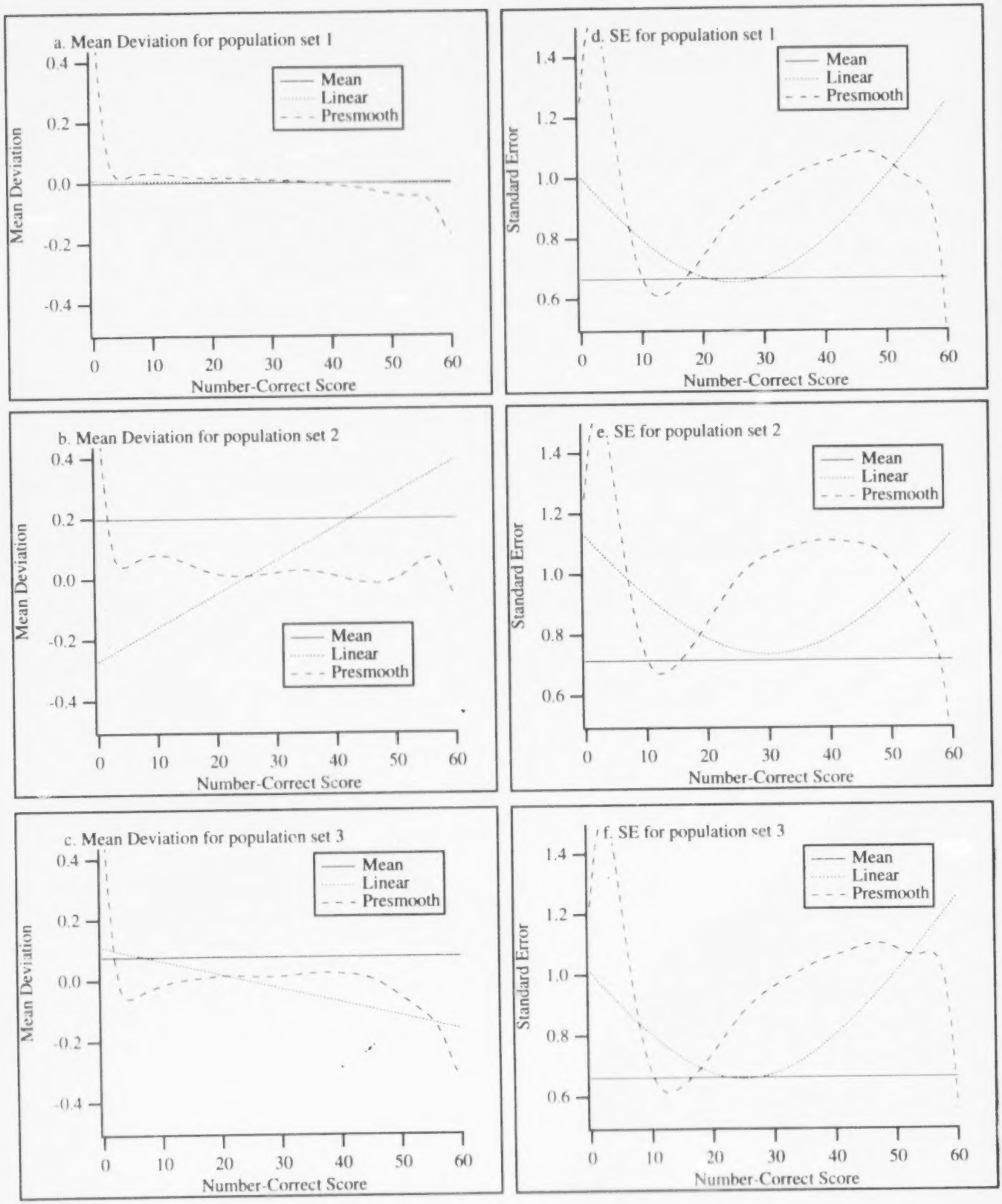

Figure 1. The mean deviation and standard error of the circular equating for different populations and equating methods.

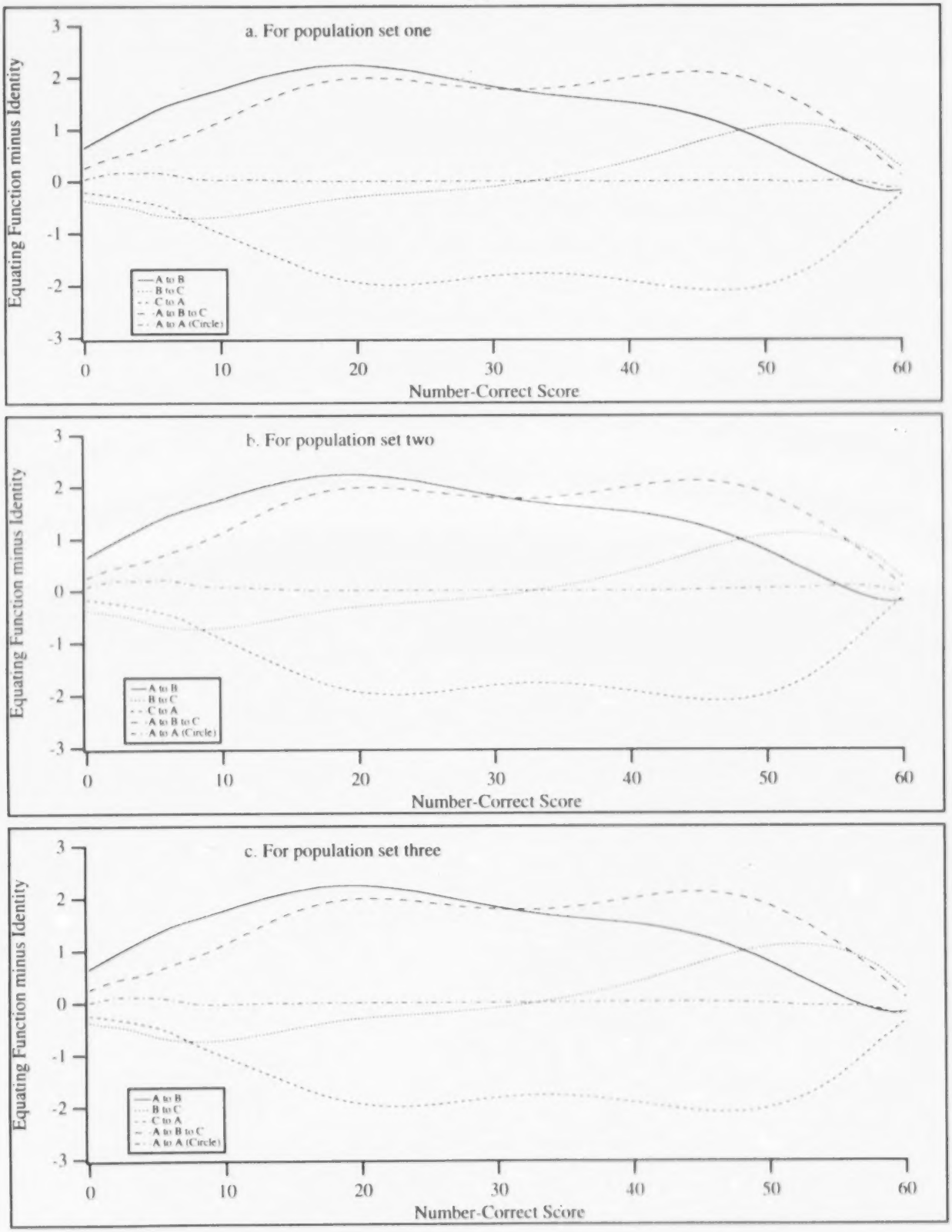

Figure 2. The true population equating functions minus identity function for the random groups design done in different years.

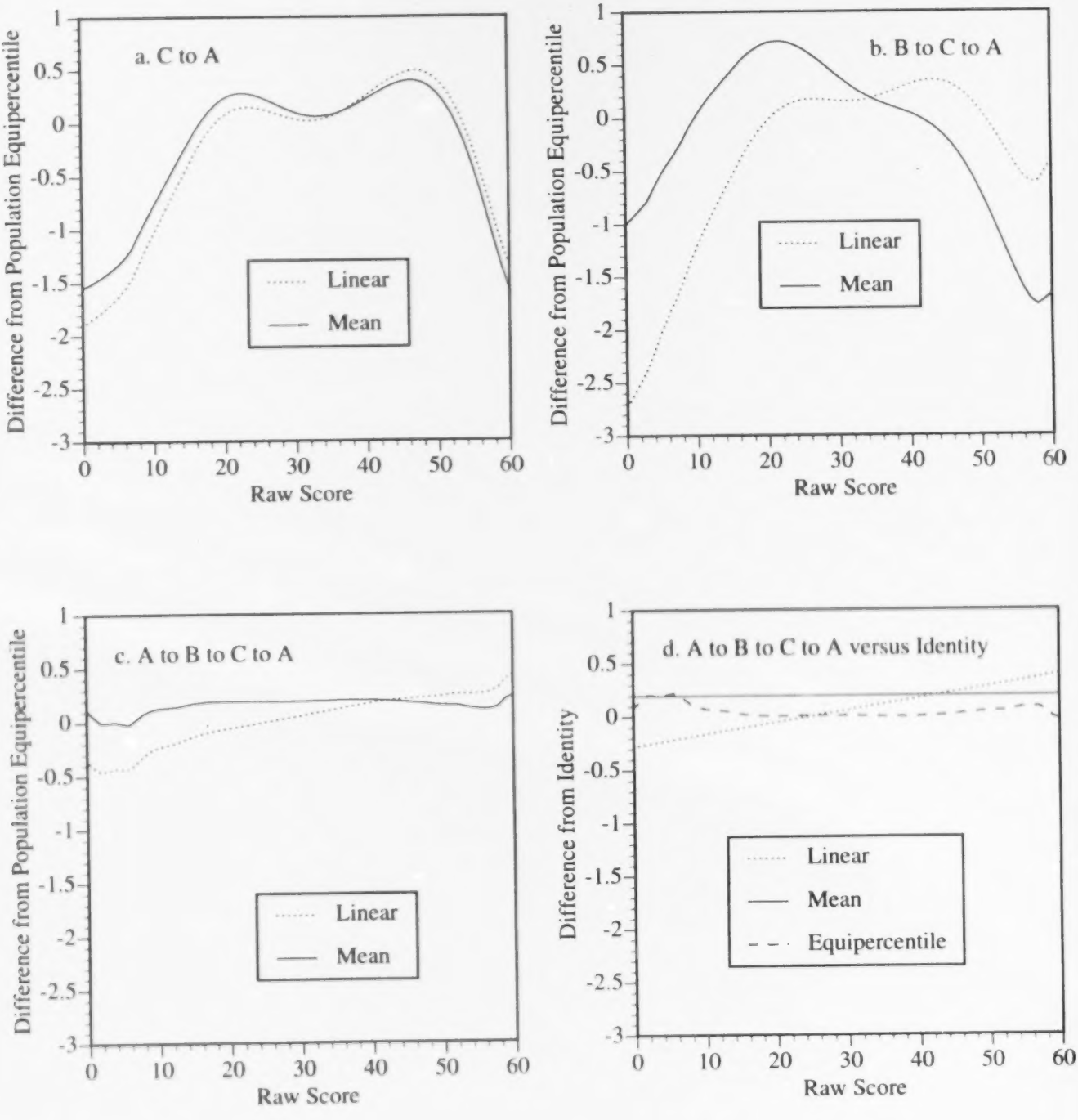

Figure 3. Population Bias of Mean and Linear Equating for the random groups design done in different years for Population Set 2.

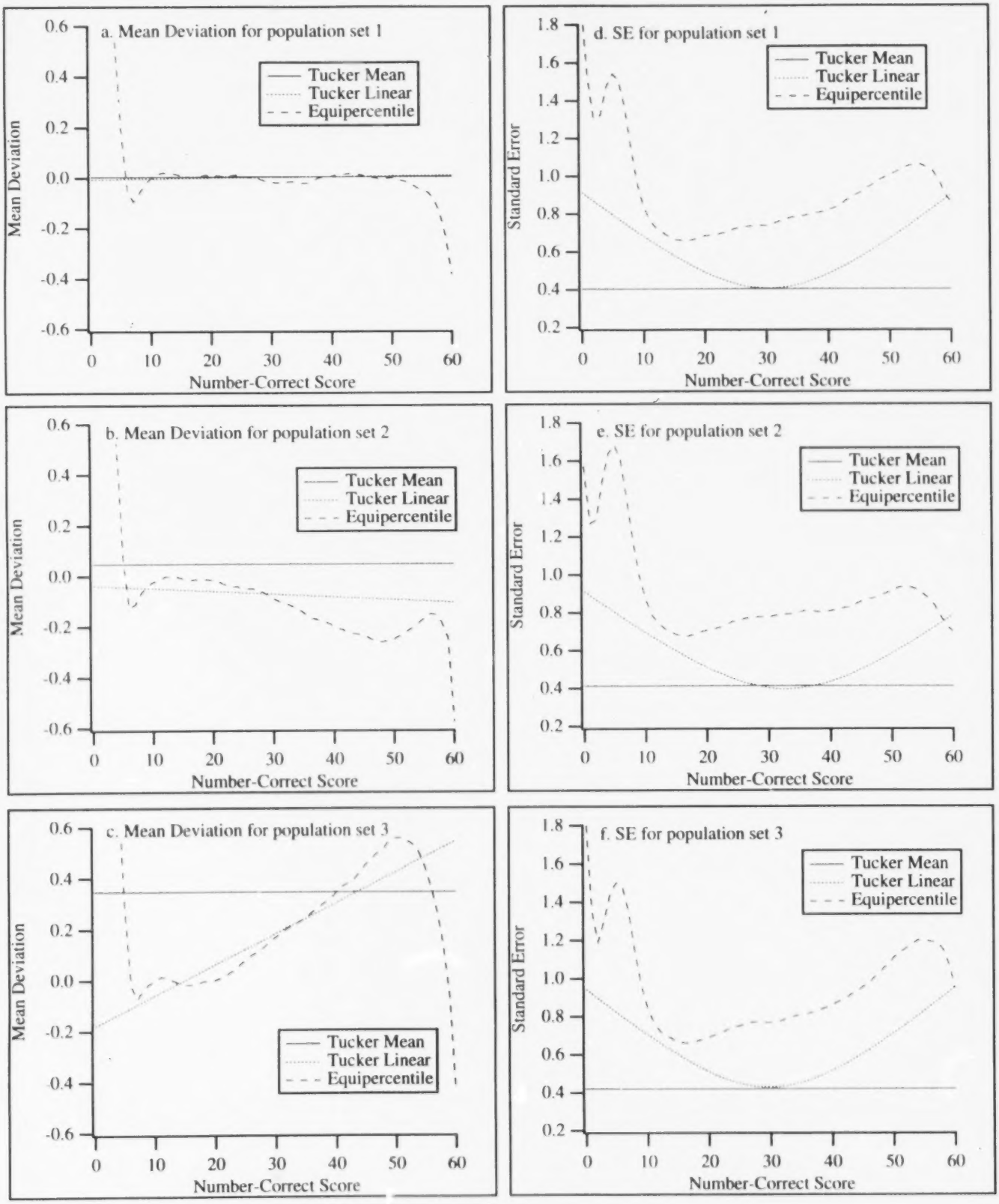
a. Mean Deviation for population set 1
b. Mean Deviation for population set 2
c. Mean Deviation for population set 3
d. SE for population set 1
e. SE for population set 2
f. SE for population set 3
Mean Deviation
Standard Error
Number-Correct Score
Tucker Mean
Tucker Linear
Equipercentile

Figure 4. The mean deviation and standard error of the circular equating for the common item nonequivalent grous design.